**GitHub Data Dive - Project Report**

**1. Introduction**

In today's rapidly evolving software development landscape, GitHub serves as a pivotal platform for collaboration and innovation in the open-source community. The GitHub Data Dive project was undertaken to extract, analyze, and visualize data from GitHub repositories focused on trending topics in the data world. This analysis aims to uncover patterns in repository characteristics, popularity, technology usage, and licensing trends, providing insights into the open-source ecosystem's dynamics.

**2. Data Extraction**

The data extraction process involved the following steps to gather comprehensive information on trending topics within GitHub repositories:

* **GitHub API Authentication:** A GitHub API token was created to securely authenticate and interact with the GitHub API.
* **Trending Topics Identification:** The project focused on the following 10 trending topics in the data world:
* Machine Learning
* Data Science
* Artificial Intelligence
* Deep Learning
* Big Data
* Data Visualization
* Natural Language Processing
* Data Engineering
* Cloud Computing
* Data Analytics
* **Data Collection:** Extracted data from 10,000 repositories (1,000 repositories per topic) using the GitHub API to ensure a diverse dataset.
* **Data Fields:** The following attributes were gathered for each repository:
* Topic
* Repository Name
* Owner
* Description
* URL
* Programming Language
* Creation Date
* Last Updated Date
* Number of Stars
* Number of Forks
* Number of Open Issues
* License Type

**3. Data Cleaning and Preprocessing**

The data cleaning process ensured that the dataset was consistent and prepared for accurate analysis. Key steps included:

* **Standardizing Programming Languages:** Replaced occurrences of 'Jupyter Notebook' in the 'Programming Language' column with 'Python', reflecting the fact that Jupyter Notebooks primarily use Python.
* **Handling Missing Values:** Addressed missing data in critical columns:  
  - Filled missing values in the 'Programming Language' column (approximately 13%) by grouping data by topic and using the mode of the programming language for each topic.  
  - Filled missing values in the 'Description' column (around 1%) with a generic placeholder where appropriate.
* **Date Formatting:** Converted timestamps from 'YYYY-MM-DDTHH:MM' format to '%Y-%m-%d %H:%M:%S' for both 'Creation Date' and 'Last Updated Date' columns to standardize date representations.

**4. Data Storage**

The cleaned data was stored efficiently in a MySQL database to facilitate easy access and management for analysis:

* **Database Setup:** Created a new database with a structured schema optimized for storing GitHub repository information.
* **Data Insertion:** Loaded the cleaned dataset into the SQL database, ensuring seamless integration for analysis and query execution.

**5. Exploratory Data Analysis (EDA)**

The exploratory analysis was performed using various visualizations to identify trends and patterns in the GitHub repository data. Key insights from this analysis include:

* **Repository Trends Over Time:** The analysis of yearly trends revealed a significant increase in the number of stars, forks, and open issues, peaking in the years 2017 and 2018.
* **Programming Language Usage:** A heatmap visualized the top 10 programming languages usage by topic. Python emerged as the dominant language across most topics.
* **License Distribution:** The license analysis showed that the MIT License was the most widely adopted, followed by repositories with no specified license and the Apache License 2.0.
* **Repository Popularity:** The analysis of repository popularity metrics identified that projects related to data science and machine learning have the highest engagement.

**6. Visualizations and Key Findings**

The following key charts and visualizations were used to represent the data insights:

* **Yearly Trend Analysis:** Line charts illustrating the growth in the number of stars, forks, and open issues over time.
* **Programming Language Heatmap:** A heatmap displaying the usage of top programming languages across different trending topics.
* **License Distribution Chart:** A bar chart showing the distribution of licenses among repositories.
* **Topic-Based Analysis:** Charts analyzing repository metrics for each trending topic.

**7. Project Outcomes and Insights**

The GitHub Data Dive project led to the following conclusions:

* **Programming Language Trends:** Python remains the most preferred language across data-related topics.
* **Open Source Licensing**: The widespread use of permissive licenses like MIT and Apache indicates a preference for code reuse.
* **Community Engagement**: Machine learning and data science repositories show the highest levels of community engagement.
* **Data-Driven Insights:** This analysis provides valuable insights into the open-source landscape.